UNIVERSITI PUTRA MALAYSIA

OPTIMIZATION OF MOTION COMPENSATED BLOCK-BASED DCT VIDEO COMPRESSION FOR SOFTWARE IMPLEMENTATION

CHEN SOONG DER

OPTIMIZATION OF MOTION COMPENSATED BLOCK-BASED DCT VIDEO COMPRESSION FOR SOFTWARE IMPLEMENTATION

By

CHEN SOONG DER

Thesis Submitted in Fulfilment of the Requirements for the Degree of Master of Science in the Faculty of Engineering Universiti Putra Malaysia

April 2000

Abstract of the thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Master of Science

OPTIMIZATION OF MOTION COMPENSATED BLOCK-BASED DCT VIDEO COMPRESSION FOR SOFTWARE IMPLEMENTATION

By

CHEN SOONG DER

April 2000

Chairman : Abd. Rahman Ramli, Ph.D

Faculty : Engineering

Internet has inspired the rapid development of wide range of network application such as Tele-Conferencing, Distance-Learning, Tele-Medicine etc, in which real time video delivery plays an important role. Due to the nature of video, which is large in size, video compression is essential in determining the practical implementation of network video application. Current video compression standards such as MPEG-1, MPEG2, H.261 and H.263 employ motion compensated DCT (Discrete Cosine Transform) block-based compression schemes and offers good compression ratio. However, it requires high processing power in order to achieve real-time processing. Therefore, optimizations are desirable especially when software implementation is preferred for its flexibility as compared to hardware implementation.

This thesis focuses on ways to improve the existing solutions in the algorithmic and implementation aspects. For the algorithmic aspect, the basic principles of motion compensated DCT block-based compression scheme was studied. Then, various optimized algorithms for the two core processes in the compression, DCT and motion estimation, were reviewed and analyzed. For the

implementation aspect, software-driven media processing was studied. A popular software-driven media processing's technology - MMX™ was studied for its application in 2-D 8x8 DCT.

The above studies and reviews provide two proposals for improvements. The first proposal is a method based on the energy preservation theorem to be applied in the H.263 video compression standard to detect frequent All-Zero-AC coefficient blocks. When such a block was detected, some of the standard processing steps may be skipped and some computation may be saved. The proposed new algorithm was evaluated and the results indicate that it was practical in low bit rate environments targeted by H.263 as no negative speed gain was observed for the full range of step size during the evaluation

Existing MMX implementation of 2-D 8x8 IDCT with uniform 16-bit precision can hardly pass the IEEE standard compliance test, which serve to prevent Inverse DCT mismatch that can cause serious distortion in decoded video. Therefore, the second proposal suggests a standard compliance implementation with mixed 32/16-bit precision and rounding. The mixed 32/16-bit design has the capability to absorb the extra operations incurred by 32-bit operation through eliminating the need for matrix transposition. The proposed implementation's precision was further improved by rounding before it could pass the entire test. Result shows that the proposed implementation needed only small increment (<10%) in overall operations in order to be standard compliance.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PENGOPTIMUMAN KE ATAS PEMAMPATAN VIDEO YANG MENGGUNAKAN SKEMA IMBUHAN-ALIHAN DENGAN TRANSFORMASI KOSIN DISKRIT BERBLOK UNTUK TUJUAN PEMBANGUNAN SECARA PERISIAN

Oleh

CHEN SOONG DER

April 2000

Pengerusi : Abd. Rahman Ramli, Ph.D

Fakulti : Kejuruteraan

Internet telah merangsangkan pembangunan yang pesat bagi pelbagai aplikasi rangkaian seperti Tele-Mesyuarat, Pembelajaran Jarak Jauh, Tele-Perubatan dan lain-lain yang mana, penghantaran video secara masa nyata memainkan peranan penting. Akibat daripada saiz maklumat video yang lazimnya begitu besar, pemampatan video telah menjadi satu keperluan dalam menentukan kejayaan sesuatu aplikasi video dalam rangkaian. Piawaian pemampatan video yang sedia ada seperti MPEG-1, MPEG-2, H.261 dan H.263 yang menggunakan skema pemampatan video yang menggunakan skema imbuhan-alihan dengan transformasi kosin diskrit berblok berupaya untuk menghasilkan nisbah pemampatan yang tinggi. Walau bagaimanapun, skema tersebut memerlukan kuasa pemprosesan yang tinggi untuk mencapai pemprosesan masa nyata. Oleh itu, penyesuaian adalah amat diperlukan terutama apabila pembangunan secara perisian lebih digemari kerana ia lebih timbal-balik berbanding dengan pembangunan secara perkakasan.

Dalam tesis ini, usaha ditumpukan dalam memperbaiki penyelesaian yang sedia ada dalam kedua-dua aspek algoritma dan aspek pembangunan. Bagi aspek algoritma, prinsip-prinsip asas bagi skema pemampatan berasaskan imbangan-alihan

dengan transformasi kosin diskrit berblok telah dikaji. Kemudian, pelbagai algorithma yang telah diubahsuaikan bagi dua proses utama dalam pemampatan, iaitu: transformasi kosin diskrit dan penganggaran alihan telah disorot and dianalisis. Bagi aspek pembangunan, pemprosesan media dengan pacuan perisian telah dikaji. Satu teknologi yang popular dalam pemprosesan media dengan pacuan perisian - MMX™ telah dikaji terhadap aplikasinya yang melaksanakan transformasi kosin diskrit dua dimensi 8x8.

Kajian dan sorotan ini telah membawa hasil kepada dua cadangan perbaikan. Yang pertama mencadangkan satu cara yang berasaskan teorem pemeliharaan tenaga yang bakal digunakan dalam piawaian pemampatan video H.263 untuk mengesan blok yang hanya perlu diwakili oleh satu nilai purata. Apabila blok sebegini dikesan, sebahagian daripada langkah pemprosesan boleh dikecualikan. Algoritma baru yang dicadangkan telah diuji dan keputusannya menunjukkan bahawa ia adalah berguna dalam keadaan kadar bit rendah seperti yang disasarkan oleh H.263. Ini adalah kerana tiada perolehan kelajuan negatif diperhatikan semasa proses penilaian untuk kesemua nilai langkah.

Rutin bagi proses transformasi pembalikan kosin diskrit dua dimensi 8x8 yang menggunakan MMX dengan kejituan seragam 16 bit tidak berupaya memenuhi ujian penyesuaian piawaian IEEE yang bertujuan untuk mengelakkan berlakunya ketidak-serasian transformasi tersebut yang akan memberi kesan buruk terhadap video yang dinyahkodkan. Oleh itu, cadangan kedua mencadangkan suatu skema pembangunan bagi rutin tersebut yang memenuhi keperluan ujian penyesuaian piawaian dengan menggunakan kejituan campuran 32/16-bit dan pembulatan. Reka bentuk kejituan campuran 32/16-bit mempunyai keupayaan untuk menyerap operasi tambahan yang disebabkan oleh penggunaan operasi 32-bit dengan mengelakkan

keperluan untuk permindahan matriks. Kejituan bagi skema pembangunan yang dicadangkan telah diperbaiki lagi dengan menggunakan pembulatan sebelum ia dapat memenuhi kesemua permintaan ujian penyesuaian. Keputusan telah menunjukkan bahawa skema pembangunan yang dicadangkan hanya memerlukan peningkatan yang sidikit (<10%) ke atas keseluruhan operasi untuk memenuhi syarat penyesuaian piawaian

ACKNOWLEDGEMENTS

I am indebted to several people who have directly or indirectly contributed to this dissertation work. The knowledge and experiences I have gained with them cannot be measured in words. Neither can the appreciation I have for them. The following list, by no means exhaustive, is an attempt to acknowledge at least a few of these people.

First and foremost, it is my pleasure to express my deepest gratitude and appreciation to Dr. Abd. Rahman Ramli, the chairman of my supervisory committee for his guidance and inspiration during the research period. I would also like to express my sincere thanks to Prof. M.R Mukerjee for his kind supervision and suggestion to carry out my works properly. Not to forget also Prof. Dr. M. N. Faruqi who has been my examiner despite of his tight schedule, Dr. V. Prakash who has shared much of his technical knowledge and Assoc. Prof. Dr. Borhanuddin Mohd. Ali who has generously allowed me to access his lab facilities.

I also wish to take this opportunity to express my sincere gratitude to all my fellow friends in Multimedia Research Group as well as ATM & Broadband Research Group for their help, friendships and insightful discussion. Not to forget also all my friends who are always with me when I was upset or emotionally down. Their unconditional supports and encouragement have enabled me to go through many hard times in life. I am particularly indebted to Low Wai Yan, Chee Boon Kok, Ang Chee Huei, Beh Kok Siang, Choong K.N and Ali Mohd. Abdelrahman.

Above all, I owe my humble achievements to my family back home. My parent deserves the credit for my degree. They have provided mental support and encouragement for my higher studies and taught me to take pride in my work.

I certify that an Examination Committee met on 26th April 2000 to conduct the final examination of Chen Soong Der on his Master of Science thesis entitled "Optimization of Motion Compensated Block-based DCT Video Compression for Software Implementation" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

MD. MAHMUD HASAN, Ph.D.
Faculty of Engineering,
Universiti Putra Malaysia.
(Chairman)

ABD. RAHMAN RAMLI, Ph.D.
Faculty of Engineering,
Universiti Putra Malaysia.
(Member)

VEERARAGHAVAN PRAKASH, Ph.D.
Faculty of Engineering,
Universiti Putra Malaysia.
(Member)

M. N. FARUQUI, Ph.D.
Professor
Faculty of Engineering,
Universiti Putra Malaysia.
(Member)

MOHD. GHAZALI MOHAYIDIN, Ph.D.
Professor/Deputy Dean of Graduate School,

Date: 03 MAY 2000

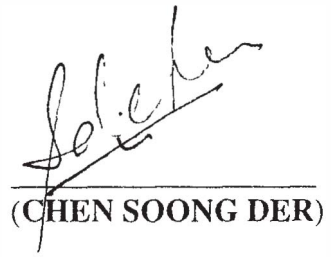
This thesis submitted to the Senate of Universiti Putra Malaysia and was accepted as fulfilment of the requirements for the degree of Master of Science.

KAMIS AWANG, Ph.D.
Associate Professor
Dean of Graduate School,
Universiti Putra Malaysia

Date: 8 JUN 2000

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.



(CHEN SOONG DER)

Date: 2 May 2000

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBRIVATIONS

1-D	-	1 Dimension
2-D	-	2 Dimension
BPM	-	Bit-plane Matching
BPROP	-	Binary Level Matching Criterion
CCIR	-	Consultative Committee for International Radio
CIF	-	Common Interchange Format
CPU	-	Central Processing Unit
DCT	-	Discrete Cosine Transform
DFT	-	Discrete Fourier Transform
DHT	-	Discrete Hadamard Transform
DPC	-	Difference Pixel Count
DSP	-	Digital Signal Processing
DST	-	Discrete Sine Transform
FBMA	-	Full-search Block Matching Algorithm
FDCT	-	Forward Discrete Cosine Transform
fps	-	frame per second
FSS	-	Four Step Search
GOPS	-	Giga Operation Per Second
HME	-	Hierarchical Motion Estimation
IDCT	-	Inverse Discrete Cosine Transform
IEEE	-	Institute of Electrical and Electronic Engineers
ITU	-	International Telecommunication Union
JPEG	-	Joint Photographic Experts Group

Kbps	-	Kilo bits per second
KLT	-	Karhunen-Loeve Transform
MAD	-	Mean Absolute Difference
MAE	-	Mean Absolute Error
Mbps	-	Mega bits per second
MOPS	-	Mega Operation Per Second
MPEG	-	Motion Picture Expers Group
MSD	-	Mean Square Difference
MSE	-	Mean Square Error
NFSS	-	New Four Step Search
NFTSS	-	New Fast Three Step Search
NTSC	-	National Television System Committee
NTSS	-	New Three Step Search
ome	-	overall mean error
omse	-	overall mean square error
PAL	-	Phase Alternative Line
PC	-	Personal Computer
PDC	-	Pixel Difference Classification
PHODS	-	Parallel Hierarchical One-Dimensional Search
pme	-	pel mean error
pmse	-	pel mean square error
POTS	-	Plain Old Telephone Switching
ppe	-	pel peak error
QCIF	-	Quad Common Interchange Format
SAD	-	Sum Absolute Difference

CHAPTER I

INTRODUCTION

Real-Time Video Delivery Via Internet

The wide spread of Internet as the most cost effective global computer network has opened up a new era in telecommunication. It has inspired the rapid development of wide range of network applications such as tele-conferencing, distance-learning, tele-medicine etc. In most of the application stated above, real time video delivery has been the essential part.

Despite Internet's wide connectivity, its limited bandwidth raises problem in real time video delivery. Uncompressed digital video usually requires high bandwidth. The bandwidth required for transmitting an entry level video sequence such as QCIF (Quad Common Interchange Format) is approximately 9 Mbps (Mega bit per second). In contrast, the current maximum Internet connection speed for a dialup user is only 56 Kbps (Kilo bit per second) while a lease line user could hardly reach 1 Mbps. This situation has made video compression more important than ever before in determining the practical implementation of network video applications.

Available compression standards such as MPEG-1 (ISO/IEC JTC1 CD 11172, 1992), MPEG-2 (ISO/IEC JTC1 CD 13818, 1994), H.261 (ITU-T Recommendation H.261, 1993), H.263 (ITU-T Recommendation H.263, 1995) etc can offer compression ratios up to 400:1. However, compression also causes negative effects. While it may possibly reduce the video quality, the compression and decompression processes also take time and cause compression latency. In

interactive video applications such as video conferencing, compression latency must be kept minimized so that the interaction will not be disturbed.

Hardware implementation of video compression offers high processing speed but often requires high cost and also lack of flexibility. Software implementation is well known as a low cost and flexible solution but often suffers from low speed problems. Current trends are towards developing a high speed, flexible, yet low cost software driven solution. In order to achieve this goal, optimization must be carried out in both implementation and algorithmic aspects of video compression.

Research Directions

In algorithmic aspect, many researches have been carried out to optimize the algorithms used to perform two core processes in video compression standard, which are DCT (Discrete Cosine Transform) and motion estimation. Fast DCT algorithms that require much less computation compared to the native algorithm have been developed through exploiting the mathematical properties of DCT. Nevertheless, under certain circumstances, there tends to be many blocks that are consist of only DC (zero frequency) coefficient with all AC (none zero frequency) coefficients neglected (quantized to zero). This happens rather frequent in low bit rate communications where motion activity is low with respect to the imposed quantization step size. A good example is video conferencing where in most of the time, there are only small movements of speaker's head and shoulder. In such situation, the computation of complete transformation could have been a waste as DC coefficient can be easily obtained from the mean of the block. Therefore, if it is possible to detect the occurrence of such situation by some means that requires less

operation, some savings in computation could be achieved. This thesis serves to formulate an effective detection algorithm.

In implementation aspect, SIMD (Single Instruction Multiple Data) technology such as MMX™ technology (Mittal et at, 1997) found in new generation PC microprocessors has brought significant speed up to various multimedia application including video compression. For example, the MMX implementation of 2-D DCT has shown speed gain up to 3 times over the conventional C programming implementation. However, since the technology is based on fixed-point arithmetic operation, calculation accuracy has become a major issue. Existing uniform 16-bit precision implementation scheme for 2-D IDCT (Inverse Discrete Cosine Transform) has been found unable to meet the accuracy requirement of IEEE standard 1180-1990 (IEEE Standard Specifications for the Implementation of the 8x8 IDCT). Therefore, it is prone to IDCT mismatch that could cause serious distortion in decoded video sequence. Unfortunately, a direct shift to higher precision (32-bit) implementation may significantly reduce the speed gain. Therefore effort must be spend on developing a precise and high-speed MMX implementation of 2-D IDCT for software based video compression.

Thesis Objectives

The aim of this thesis is to contribute towards realizing real time video compression via software implementation. Therefore, the objectives have been set as follows:

- To develop an efficient algorithm that could detect the occurrence of All-Zero-AC block with high detection accuracy and low computation overhead.

- To develop an MMX implementation scheme for 2-D IDCT that could achieve standard requirement while maintaining the speed as much as possible.

Thesis Organization

This thesis is organized into six chapters. Chapter I gives brief introduction to real time video delivery and related issues. This chapter also defines the intended research areas, the objectives and the organization of the thesis. Chapter II presents the literature review on various existing fast/optimized compression algorithms. The MMX™ technology is also reviewed on its implementation of DCT and its precision limitation in preventing IDCT mismatch is brought up for discussion. Chapter III explains the background theories of the existing video compression standards as well as the principle of software media processing with MMX™ technology. Chapter IV describes in details the principle, design and experiment setup of two proposals for improvement: a) All-Zero-AC blocks detection using energy preservation theorem. b) Standard compliance MMX implementation of 8x8 IDCT. Chapter V discusses the evaluation results for these two proposals. The final chapter serves to conclude this thesis with highlight of thesis' contributions and discussions on future research direction.

CHAPTER II

LITERATURE REVIEW

In this chapter, extensive review is carried out on the existing optimizations in both algorithmic and implementation aspect. The first and second section reviews the existing optimized algorithms for DCT as well as motion estimation. These sections aims to established understanding on the principles the fast algorithms as well as the indication of research direction The third section explains the existing MMX implementation scheme for DCT, including its principles, details operation, limitation and topics related to precision. The last section concludes this chapter by highlighting the proposed research direction.

Fast DCT Algorithms

In this section, review will be carried out on some of the principles and techniques used in fast DCT algorithms (Cooley and Tukey, 1965), the origins of which go back to algorithm for the Fast Fourier Transform implementation of the Discrete Fourier Transform (DFT). An extensive review of fast DFT techniques published by Duhamel and Vetterli (1990) contains an interesting section on the history of this field of research. Fast DCT techniques as a separate field started with the paper by Chen et al. (1977), but the linkage between the DCT and the DFT continues to be important.

The most straightforward way to perform the DCT is to follow the theoretical equations. This will require 64 multiplication and 56 additions for each 1-D 8 point DCT. Note that the cosine terms can be combined with the constants before the computation because the cosines become discrete numbers at each position. Therefore, a full 8 x 8 DCT done in separable 1-D format (eight rows and eight columns) would require 1024 multiplication and 896 additions plus additional operations to quantize the coefficients. This is considerably high in term of computation complexity and thus, fast algorithms are desirable for practical use of DCT.

The fast DCT techniques take advantages of the symmetries in the DCT equations and exploit the properties of the transformation matrix **T**. Essentially, the matrix **T** can be factorized so that

$$T = T_1 \bullet T_2 \bullet \ldots \bullet T_k \qquad (1)$$

where each of the matrices $\mathbf{T_1 T_2 \ldots T_k}$ is sparse. Sparseness implies that most of the elements of the matrix are zero. Therefore, the transform of **X** can be expressed as

$$T \bullet X = T_1 \bullet T_2 \bullet \ldots \bullet T_k \bullet X \qquad (2)$$

if the calculations are performed in a sequential manner; compute $\mathbf{s_k = T_k \bullet X}$ first and then use the result to perform the next matrix product, $\mathbf{s_{k-1} = T_{k-1} \bullet s_k}$, then it can be shown that there is an overall reduction in the number of operations required to perform the transformation.

As an example, Ligtenberg and Vetterli (1986) have derived a fast DCT algorithm that only require 208 multiplication and 464 additions to perform a full 2-D FDCT, giving approximately 60% reduction in the total computation required. The flowgraph for this fast DCT is as shown in Figure 1. Table 1 shows the computational

complexity of some other commonly used fast DCT algorithms (Bhaskaran and Konstantinides, 1996).

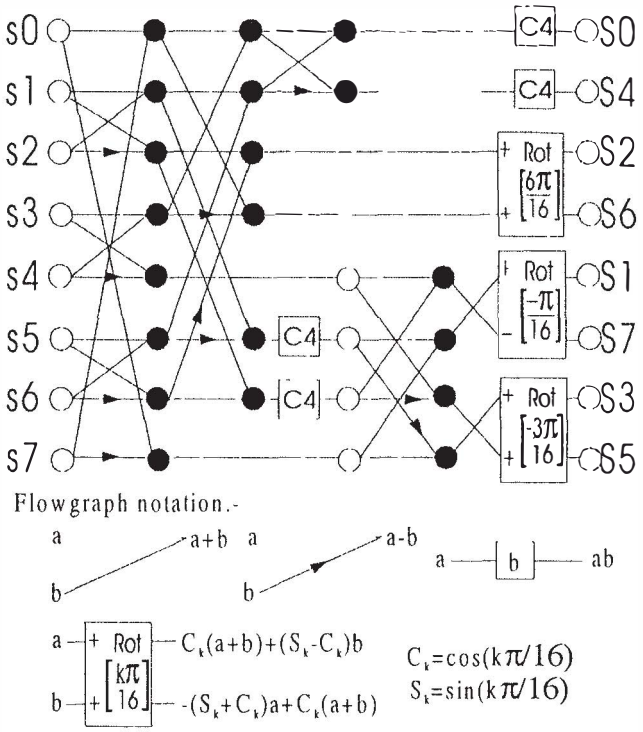


Figure 1: Flowgraph for 1-D DCT from Ligtenberg and Vetterli (1986)

Table 1: Computational complexity of various fast DCT algorithms

Fast Algorithm	Multiplications		Additions	
	1-D	2-D	1-D	2-D
1-D Chen	16	256	26	416
1-D Lee	12	192	29	464
1-D Loeffler, Ligtenberg	11	176	29	464
2-D Kamangar, Rao		128		430
2-D Cho, Lee		96		466

Note that the first three algorithms are for an eight-point 1-D DCT from which the 2-D DCT can be computed using the separability property of the DCT. The last two methods were developed exclusively for a 2-D DCT. As shown by Table 1 that a true 2-D method, such as the Cho and Lee (1991) approach, yields lower complexity than the separable extensions of the 1-D DCT. However, from an implementation viewpoint, many of the true 2-D DCT methods have several disadvantages:

1. In a software implementation, storage for up to 128 elements is needed. In a computer system, register is limited and such storage is not always feasible.
2. Data addressing is highly irregular. In software and hardware implementation, this irregularity leads to additional overhead for address calculations, which are not included in the original number of multiplication and additions.

As a result, most practical software and hardware implementations of DCT-based encoders use the row-column extension of the 1-D DCT.

There are also a number of relationships between the DFT on real inputs and the DCT. Vetterli and Nussbaumer (1984) showed that N-point DCT can be expressed in terms of the real and imaginary parts of an N-point DFT and rotations of the DFT outputs. Haralick (1976) shows that the first N coefficients of a 2N-point DFT with appropriate symmetry of input values can be used to compute an N-point DCT. His analysis later lead to a very efficient scaled DCT structure. Tseng and Miller (1978) further shows that the DCT be obtained solely from the real part of DFT when the symmetry is defined and concludes that DCT coefficients can be obtained by simple scaling of the real part of the DFT coefficients. An optimum form of 16-point DFT has been developed by Winograd (1978) and also Silverman (1977). In its most general form of 16-point DFT requires 18 multiplication and 74 additions. However, Arai et al. (1988)